isocir: An R package for Isotonic Inference for Circular data. An application in Cell Biology.

Sandra Barragán
Universidad de Valladolid

Miguel A. Fernández
Universidad de Valladolid

Cristina Rueda
Universidad de Valladolid

Shyamal Das Peddada
National Institute of
Environmental Health Sciences

Abstract

The R package **isocir** provides a set of functions for making isotonic inference for circular data. In this setting, the standard statistical methods cannot be used to make inferences due to the geometry of the circle and restrictions, estimators and hypotheses tests have to be properly defined to cope with the peculiarities of circular data. Rueda *et al.* (2009) considers the estimation problem and solves it for the appropriate circular orderings among which the isotropic order is the most suitable for applications. Fernandez *et al.* (2011) provides a methodology for dealing with isotropic testing problems.

In this paper we generalize the estimation and testing results obtained in those papers and implement the corresponding procedures in R language. Since one of the main fields where circular data are relevant is cell biology we illustrate the package with cell cycle data examples. However, we want to stress its usefulness in any context where circular data may appear.

Keywords: Circular Data, Isotropic Order, *CIRE*, Conditional Test, R package **isocir**, R.

1. Introduction

In this paper we describe the use of **isocir** package within the R statistical environment, R Development Core Team (2004), which is available from the Comprehensive R archive Network at `http://CRAN.R-project.org`. The package **isocir** provides functions that carry out the methodology for the analysis of circular data under restrictions.

Circular data arise in a wide range of contexts, such as in geography, cell biology, circadian biology, endocrinology, ornithology, etc (see Zar (1999), Mardia *et al.* (2008) or Berens (2009)). Unlike the Euclidean space, the points are wrapped around on a unit circle. That is, starting at a point "P", by traveling around the circumference of the circle one would return to the point "P". As a consequence a circle can never be linearized and hence methods developed for Euclidean space data are not applicable to circular data. The starting point "P" is said to be the pole of the circle and we use the standard convention of traveling in the counter-clockwise direction. Thus the angles are measured between $[0, 2\pi]$. General methodology for circular data can be found in the book Mardia and Jupp (2000), among others.

It is frequent that in some statistical applications, additional information is available to the researchers. In the case of Euclidean space data, the simple order restriction on population parameters is an important inequality constraint that is widely noted in practice (cf Peddada *et al.* (2007)). According to this constraint the experimenter knows a priori the relative order among all population parameters under consideration. Estimation of the population parameters under the simple order constraint is known as the isotonic regression. A popular algorithm for solving this problem is the *pool-adjacent violators algorithm* (PAVA). See Robertson *et al.* (1988) for details.

For circular parameter space, the standard notion of simple order needs to be modified to account for the fact that the parameters wrap around the circle. Furthermore, as a consequence of this characteristic of a circle, the PAVA for Euclidean space data is not directly applicable to circular data. Rueda *et al.* (2009) introduced an order restriction on a unit circle called the isotropic order. They also extended the notion of isotonic regression estimator to circular parameter space, known as the circular isotonic regression estimator (CIRE) and developed an estimation procedure which is a generalization of PAVA. In Section 2 we describe the isotropic order and CIRE in detail.

The initial motivation for developing constrained inference methods for circular data was the analysis of cell-cycle gene expression data. Since the normal cell cycle is a well orchestrated process consisting of four major phases, namely, G1, S, G2 and M, of distinct biological functions, cell biologists have long been interested in determining the phase associated with each cell cycle gene (cf Oliva *et al.* (2005), Rustici *et al.* (2004)). The current understanding is that a cell cycle gene would attain its peak expression during the phase corresponding to its biological function. For a given subset of cell cycle genes, a cell biologist may also be interested in inferring whether the relative order of peak expression among these cell cycle genes is conserved across multiple species. Until now there did not exist a formal statistical methodology for answering such questions. Recently, using the estimators derived in Rueda *et al.* (2009), Fernandez *et al.* (2011) developed a formal statistical theory and methodology for testing the isotropic order among a subset of cell cycle genes. Using this methodology, Fernandez *et al.* (2011) concluded that the isotropic order among a large subset of cell cycle genes is conserved between two species of yeasts, namely, the budding yeast and the fission yeast. They also inferred that fewer cell cycle genes were conserved between humans and fission yeast. Results such as these provide important insights into evolutionary biology since cell division is fundamental to growth and development of every organism.

Statistical methods developed in Rueda *et al.* (2009) and Fernandez *et al.* (2011) would have wide range of applications beyond the analysis of cell cycle gene expression data. For example, ornithologists may find these methods useful in their investigation of the migratory patterns and directions of birds, Cochran *et al.* (2004). An endocrinologist may find these methods useful when studying temporal patterns hormones in people treated for hormonal imbalances or researchers investigating genes controlling circadian clock. Moreover, this methodology is used in other scientific disciplines such as earth science (some feature of an earthquake), meteorology (wind directions, Bowers *et al.* (2000)), physics (orbits of planets or direction fluctuations in the atmosphere, van Doorn *et al.* (2000)), psychology (studies of mental maps

or monitoring data, Kibiak and Jonas (2007)), image analysis (the orientation of ridges on fingerprints or magnetic maps, Boles and Lohmann (2003)), medicine (the incidence of onsets of a particular disease or investigating some disease indicator, Le *et al.* (2003)), neuroscience (orientation selectivity, Maldonado *et al.* (1997)), political and social sciences (Haskey (1988)), criminology (Brunsdon and Corcoran (2005)) and many more. Motivated by the wide range of applications and the non-existance of a user friendly software, in Section 3 of this article we introduce our R based user friendly software called **isocir**. In Section 4 we illustrate the software by analyzing a cell cycle gene expression data. Concluding remarks are provided in Section 5.

2. Circular models with parameters under restrictions

2.1. Description of the order restrictions

Let φ_{ik} $\forall i = 1, \ldots, q$, $k = 1, \ldots, n_i$, be angular observations from q populations with mean directions $\phi_1, \ldots, \phi_q$. Let $\theta_1, \ldots, \theta_q$ be the sample mean directions and $r_1, \ldots, r_q$ the sample mean resultant lengths (check Mardia and Jupp (2000)).

As usually done, throughout this paper, angles are measured in the anti clockwise direction. If the pole of the circle is at zero radians and pretend that the parameters are points on the line then the usual notion of simple order would be:

$$C_{SO} = \{\phi \in [0, 2\pi]^q / 0 \leq \phi_1 \leq \phi_2 \leq \ldots \leq \phi_q \leq 2\pi\} \tag{1}$$

A problem with the above representation is that it does not acknowledge that the angle ϕ_q is "followed by" ϕ_1. There is a disconnect between the two parameters in the above representation. In many practical applications, such as in cell biology (see Rueda *et al.* (2009), and Fernandez *et al.* (2011)), such a disconnect is not meaningful. This is because, as far as the biologist is concerned there is a relative order among all q parameters. Thus the usual notion of simple order defined for parameters in the Euclidean space may not be appropriate for circular parameters.

In view of this need, Rueda *et al.* (2009) introduced the following order restriction on a circle which is called the isotropic order. Suppose $\phi_i, i = 1, 2, 3$, are three angular parameters on a unit circle. Then they are said to be in an isotropic order if ϕ_1 is "followed" by ϕ_2 which is "followed" by ϕ_3 which in turn us followed by ϕ_1. We use the notation $\phi_1 \preceq \phi_2 \preceq \phi_3 \preceq \phi_1$. More generally, the following notation is used to describe the isotropic order among the q angular parameters:

$$C_{IO} = \{\phi \in [0, 2\pi]^q / \phi_1 \preceq \phi_2 \preceq \ldots \preceq \phi_q \preceq \phi_1\} \tag{2}$$

Thus for all $i = 1...q$, ϕ_i is after ϕ_{i-1} and before ϕ_{i+1}, $\phi_0 \equiv \phi_q$ and $\phi_{q+1} \equiv \phi_1$. The isotropic order does not depend on the location of the pole and is rotation invariant.

Let $C_{SO}^{I} = \{0 \leq \phi_I \leq \phi_{I+1} \leq \ldots \leq \phi_q \leq \phi_1 \leq \ldots \leq \phi_{I-1} \leq 2\pi\}$ be the simple order starting at index I. Then the isotropic order cone is:

$$C_{IO} = \bigcup_{1 \leq I \leq q} C_{SO}^{I} \tag{3}$$

From a practical point of view it is also interesting to consider the following generalization of the isotropic order. It allows taking into account possible partial order relations among groups of parameters.

$$C_{GIO} = \left\{ \phi \in [0, 2\pi]^q : \begin{Bmatrix} \phi_{11} \\ \phi_{12} \\ \vdots \\ \phi_{1l_1} \end{Bmatrix} \preceq \begin{Bmatrix} \phi_{21} \\ \phi_{22} \\ \vdots \\ \phi_{2l_2} \end{Bmatrix} \preceq \ldots \preceq \begin{Bmatrix} \phi_{L1} \\ \phi_{L2} \\ \vdots \\ \phi_{Ll_L} \end{Bmatrix} \preceq \begin{Bmatrix} \phi_{11} \\ \phi_{12} \\ \vdots \\ \phi_{1l_1} \end{Bmatrix} \right\}, \tag{4}$$

where L is the number of groups in the order, l_j is the number of angular parameters in the j level and q is the total number of parameters $q = \sum_{j=1}^{L} l_j$.

In this case, we assume that each of the parameters in group j, $\{\phi_{j1}, \ldots, \phi_{jl_j}\}$ follow the ones in group $j-1$ and are followed by the ones in group $j+1$ but we do not assume any order among the parameters inside each group. This generalized isotropic order plays an important role in cell biology when a biologist is investigating a large number of cell cycle-genes. In such situations, it may be difficult for a biologist to ascertain the exact order among all cell-cycle genes under consideration.

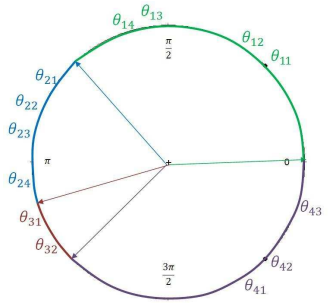

Figure 1: Graphical example

However, the biologist may be able order groups of genes based on their known biological functions. In such situations the proposed generalized isotropic order can be a natural constraint on the parameter space.

Figure 1 is a graphical example where data follow this general isotropic order. Each group is a set of parameters. In this example of the general isotropic order, there are 13 parameters divided in four groups, $L = 4$. There is a different number of elements in each group: $l_1 = 4$; $l_2 = 4$; $l_3 = 2$; $l_4 = 3$, so $q = \sum_{l}^{L} l_j = 13$.

2.2. CIRE (Circular Isotonic Regression Estimator)

Solutions to a wide class of order restricted estimation problems in the Euclidean space can be obtained using isotonic regression. The circular version of this procedure is called the Circular Isotonic Regression Estimator (CIRE) and is described in Rueda *et al.* (2009). The CIRE of ϕ, under the constraint $\phi \in C$, is given by:

$$\widetilde{\theta} = \arg\min_{\alpha \in C} SCE(\alpha, \theta), \tag{5}$$

where SCE is the sum of circular error, which is a circle analog of sum of squares error (SSE) used for Euclidean data and is defined as follows:

$$SCE(\theta, \phi) = \sum_{i=1}^{q} r_i(1 - cos(\theta_i - \phi_i)) \tag{6}$$

Although, on the face, the above minimization problem appears to be simple, it is a challenging problem as demonstrated in Rueda *et al.* (2009)). In the case of simple order in Euclidean space it is common to use the *pool adjacent violators algorithm* (PAVA) (cf Robertson *et al.* (1988)) to derive the isotonic regression estimator which minimizes the SSE. However, in the case of circle the PAVA cannot be used to derive CIRE. To apply PAVA it is essential that the Cauchy mean value property is fulfilled (cf. Robertson and Wright (1980)). According to this property, the mean value of two real numbers is strictly between the two numbers. This is not true in the case of circle. For this reason, Rueda *et al.* (2009)) developed an alternate computationally simple algorithm to derive CIRE. For more details regarding the existence, uniqueness and other properties of CIRE we refer the reader to Rueda *et al.* (2009)). The algorithm is implemented in R within the function *CIREi* in the package **isocir** which is illustrated in Section 3.

2.3. Inferences in von Mises models

From the point of view of statistical inference, perhaps the most useful and popular distribution on the circle is the von Mises distribution. This distribution is analogous to the Normal distribution on a real line.

Let $\theta_1, \ldots \theta_q$ be sample mean directions of the q independent populations. From now on, we use the notation $\theta_i \rightsquigarrow VM(\phi_i, \kappa)$, where ϕ_i is the mean direction of the population i and κ is the common concentration parameter of the von Mises distributions. The probability density function is given by:

$$f(\phi, \mu, k) = \frac{1}{2\pi I_0(k)} e^{\kappa \cos(\phi - \mu)}, \tag{7}$$

for $0 \leq \phi \leq 2\pi$, with $0 \leq \mu \leq 2\pi$ and $\kappa \geq 0$. Where I_0 denotes the modified Bessel function of the first kind and order 0.

Under this probability model Rueda *et al.* (2009) show that CIRE is the Restricted Maximum Likelihood Estimator (RMLE).

Recently, motivated by a problem in cell biology, Fernandez *et al.* (2011) tested the hypotheses given below for the isotropic order cone (C_{IO}). In the present paper, we extend those procedures to the case of testing the general isotropic order cone (C_{GIO}).

H_0 : The parameters ϕ_i $i = 1 \ldots q$ follow a known (general) isotropic order (i.e. $\phi \in C$ where C is the order cone).

H_1 : H_0 is not true (i.e. $\phi \notin C$).

If κ is known, the likelihood ratio statistic for these hypotheses is:

$$T = 2\kappa SCE(\theta, \widetilde{\theta}), \tag{8}$$

where $\widetilde{\theta}$ is the CIRE computed under the isotropic order set in H_0.
Due to computational issues related to the derivation of the critical values of the likelihood ratio test Fernandez *et al.* (2011) proposed the following asymptotic α level conditional test. This test is a modification of the likelihood ratio test which benefits from increased statistical power for interesting alternatives and is computationally very simple:

$$\text{CT: } H_0 \text{ is rejected whenever } T \geq c(m), \tag{9}$$

where m is the number of level sets for $\widetilde{\theta}$. As the asymptotic distribution of T, when κ is known, is χ^2_{q-m}, then $c(m)$ is chosen so that

$$pr(\chi^2_{q-m} \geq c(m)) = \frac{\alpha}{1 - pr_{\phi^0}(C)}, \tag{10}$$

where $pr_{\phi^0}(C)$ is the probability of the order cone (C) under the equality of the parameters, so $pr_{\phi^0}(C_{IO}) = \frac{1}{(q-1)!}$ if we test the isotropic order or $pr_{\phi^0}(C_{GIO}) = \frac{l_1!\cdots l_L!}{(q-1)!}$ if we test the general isotropic order. Note that as $T = 0$ under H_0 and we are using a conditional test, the level has to be adjusted using that probability.

If κ is unknown then it can be estimated if we have replicated data and the test statistic T can be accordingly modified as:

$$T = \frac{2\widehat{\kappa} SCE(\theta, \widetilde{\theta})}{q}, \tag{11}$$

whose asymptotic distribution under the null hypothesis is $F_{q-m,q-1}$. Thus $c(m)$ is chosen so that

$$pr(F_{q-m,q-1} \geq c(m)) = \frac{\alpha}{1 - pr_{\phi^0}(C)}, \tag{12}$$

where, as in the previous case, $pr_{\phi^0}(C_{IO}) = \frac{1}{(q-1)!}$ or $pr_{\phi^0}(C_{GIO}) = \frac{l_1!\cdots l_L!}{(q-1)!}$ depending on the null hypothesis.

These results are proved for the isotropic order (C_{IO}) in the supplementary material of Fernandez *et al.* (2011) for moderate and large values of q. Similar proofs can be obtained in the case of testing the general isotropic order (C_{GIO}).

Moreover, notice that the p-value of this test may serve as a useful goodness of fit criterion when comparing two or more plausible isotropic orders among a set of parameters. Smaller p-values, suggest that the estimations are closer to the presumed isotropic order. Thus the statistical methodology developed in Fernandez *et al.* (2011) can be used not only for testing relative order among the parameters, but it can be also useful for selecting "best fitting" isotropic order among several candidate isotropic orders for the biologist to choose form.

These tests are implemented in the function *CTi* in the R package **isocir**. In Section 4 we illustrate these tests using cell cycle gene expression data.

3. Package isocir

In the following we shall first briefly describe various R packages for isotonic regression and analysis of circular data. We shall then describe the structure of the proposed package **isocir** and demonstrate how to use it with the help of some examples.

3.1. Related R packages

Since isotonic regression is a well-known and widely used technique there are many packages in R, for performing isotonic regression, such as:

- **isotone** (de Leeuw *et al.* (2011)): Active set and generalized PAVA for isotone optimization.
- **Iso** (Turner (2009)): Functions to perform isotonic regression.
- **bisoreg** (Curtis (2010)): Bayesian Isotonic Regression with Bernstein Polynomials.
- **ordMonReg** (Balabdaoui *et al.* (2009)): Compute least squares estimates of one bounded or two ordered isotonic regression curves.
- **OrdFacReg** (Rufibach (2010)): Least squares, logistic, and Cox-regression with ordered predictors.

Similarly, there are several packages in R for analyzing circular data, such as:

- **CircStats** (Agostinelli (2009)): The implementations of the Circular Statistics from "Topics in circular Statistics" Jammalamadaka and SenGupta (2001).
- **circular** (Lund and Agostinelli (2010)): Another package with Circular Statistics from the same book, Jammalamadaka and SenGupta (2001).
- **CircSpatial** (Morphet (2009)):This package is a collection of functions for color continuous high resolution images of circular spatial data, circular kriging, and simulation of circular random fields.

Motivated by the recent interest in applications and the development of constrained inference for circular data, we introduce the software package **isocir**. The name comes from the fact that it allows making **iso**tonic inference for **cir**cular data. Our package is closely related to: **circular** (see Lund and Agostinelli (2010)) and **combinat** (see Chasalow (2010)). These packages should be installed in the computer before loading **isocir**.

3.2. Package structure

Functions used in the package are summarized in Table 1 and are briefly described below.

Functions	Arguments	Description
cirmean	(data)	circular mean
cirSCE	(arg1, arg2, mrl)	Sum of Circular Error
CIREi	(data, levels, isotropic, graphic, stack)	Circular Isotonic Regression Estimator
mrl	(data)	mean resultant length
cirkappa	(data)	kappa estimation
CTi	(data, levels, kappa)	Conditional Test

Table 1: Summary of the components of **isocir**

- **cirmean**: This function computes the circular mean as defined in Mardia and Jupp (2000).

$$\overline{\theta} = \begin{cases} \arctan\left(\frac{\overline{S}}{\overline{C}}\right) & \text{if } \overline{S} > 0, \overline{C} > 0 \\ \arctan\left(\frac{\overline{S}}{\overline{C}}\right) + \pi & \text{if } \overline{C} < 0 \\ \arctan\left(\frac{\overline{S}}{\overline{C}}\right) + 2\pi & \text{if } \overline{S} < 0, \overline{C} > 0 \end{cases} \tag{13}$$

  where $\overline{S} = \sum_{i=1}^{q} \sin\theta_i$ and $\overline{C} = \sum_{i=1}^{q} \cos\theta_i$.

- **cirSCE**: Calculates the circular error between two q dimensional points on the circle as defined in (6). In case data with replications, the mean resultant lengths (r_i) have to be introduced in the argument `mrl`. Otherwise, by default it is assumed to be 1.

- **CIREi**: For a given user specified general isotropic order (4), this function computes CIRE (5) using the algorithm developed in Rueda et al. (2009).

- **mrl**: Calculates the **m**ean **r**esultant **l**ength as defined in Mardia et al. (2000).

- **cirkappa**: Calculates the estimation of the concentration parameter of a von Mises distribution. It is necessary to have replications in the data. The argument is a matrix where each row is an individual and each replications appears in a column.

- **CTi**: Performs the conditional test and computes the corresponding p-value for the following hypotheses:

  H_0 : The phase angles ϕ_i $i = 1 \ldots q$ follow a known (general) isotropic order.

  H_1 : H_0 is not true

  The test statistic is:

$$T = \begin{cases} 2\kappa SCE(\theta, \widetilde{\theta}) & \overset{aprox.}{\rightsquigarrow} \chi^2_{q-m} & \text{if } \kappa \text{ can be assumed as known} \\ \frac{2\widehat{\kappa} SCE(\theta, \widetilde{\theta})}{q} & \overset{aprox.}{\rightsquigarrow} F_{q-m,q-1} & \text{if } \kappa \text{ is unknown (replications needed)} , \end{cases} \tag{14}$$

  where $\widetilde{\theta}$ is the CIRE obtained internally with *CIREi*, m is the number of level sets for $\widetilde{\theta}$ and $\widehat{\kappa}$ is the estimation of κ obtained when necessary with *cirkappa*. Now, if t^* is the

value of T for the data, the p-value of the test is:

$$p{-}value = \begin{cases} pr(\chi^2_{q-m} \geq t^*)[1 - pr_{\phi^0}(C)] & \text{if } \kappa \text{ can be assumed as known} \\ pr(F_{q-m,q-1} \geq t^*)[1 - pr_{\phi^0}(C)] & \text{if } \kappa \text{ is unknown (replications needed),} \end{cases} \tag{15}$$

where $pr_{\phi^0}(C)$ is the probability of the order cone (C) under the equality of the parameters.

3.3. How isocir works

We illustrate isocir by describing the two main components of the package, namely, *CIREi* for obtaining the circular isotonic regression estimator, and *CTi* for performing conditional test explained above.

CIREi

Arguments	*Values*
`data`	matrix with the data
`levels`	the levels of the order
`isotropic`	`=TRUE`(by default) / `=FALSE`
`graphic`	`=FALSE`(by default) / `=TRUE`
`stack`	`=TRUE`(by default) / `=FALSE`

Table 2: Arguments of the *CIREi* function

In this section we describe each argument of the function *CIREi*. The characteristics of these arguments are described in Table 2. The input variable `data` consists of all the input angles θ_{jp} grouped and ordered according to the desired order.

- **Example 1**

In this example, we assume the following order for the parameters:

$$\left\{ \begin{array}{c} \phi_{11} \\ \phi_{12} \\ \phi_{13} \end{array} \right\} \preceq \left\{ \begin{array}{c} \phi_{21} \\ \phi_{22} \end{array} \right\} \preceq \{\phi_{31}\} \preceq \left\{ \begin{array}{c} \phi_{41} \\ \phi_{42} \end{array} \right\} \preceq \left\{ \begin{array}{c} \phi_{11} \\ \phi_{12} \\ \phi_{13} \end{array} \right\} \tag{16}$$

Suppose that θ_{jp} $\forall j = 1, \ldots, L$, $p = 1, \ldots, l_j$, the sample mean directions corresponding to the parameters ϕ_{jp}, are:

$\theta_{11} = 0.025$; $\theta_{12} = 1.475$; $\theta_{13} = 3.274$;
$\theta_{21} = 5.518$; $\theta_{22} = 2.859$;
$\theta_{31} = 5.387$;
$\theta_{41} = 4.179$; $\theta_{42} = 1.962$.

These data are in the example set of random circular data in our package and they can be used by calling as below:

```
> data(cirdata)
```

The format of the data is a matrix and the levels of the order are defined as follows:

```
> data4 <- matrix(cirdata, ncol = 1)
> orderLevels <- c(1, 1, 1, 2, 2, 3, 4, 4)
```

The isotropic order is considered by default (i.e. `isotropic = TRUE`) but the algorithm can obtain the CIRE under the simple order by setting `isotropic = FALSE`. The result of the function is a list with three elements:

`$cirmeans` is a list with the circular means with the form set by levels.

`$SCE` is the value of the Sum of Circular Error between the data and the CIRE.

`$CIRE` is a list with the CIRE with the form set by levels.

CIRE estimates for the above example with `isotropic=TRUE` are:

```
> CIREi(data = data4, levels = orderLevels)
```

Thus,

$$\left\{ \begin{array}{l} \tilde{\theta}_{11} = 0.9939 \\ \tilde{\theta}_{12} = 1.4756 \\ \tilde{\theta}_{13} = 3.0665 \end{array} \right\} \preceq \left\{ \begin{array}{l} \tilde{\theta}_{21} = 5.0567 \\ \tilde{\theta}_{22} = 3.0665 \end{array} \right\} \preceq \{\tilde{\theta}_{31} = 5.0567\} \preceq \left\{ \begin{array}{l} \tilde{\theta}_{41} = 5.0567 \\ \tilde{\theta}_{42} = 0.9939 \end{array} \right\} \tag{17}$$

Results may be displayed graphically by setting `graphic = TRUE`. When done so, two plots are produced, one for the unrestricted estimates and the other for CIRE. Graphs are not obtained if the variable `graphic = FALSE` which is the default value. Additional arguments can be introduced to change the options of the plots, such as `stack` which is `TRUE` by default to see points with the same value separately, otherwise that points would be overlapped.

CTi

As stated earlier, in this section we assume that the sample mean directions θ_i are distributed according to independent von Mises distribution $VM(\phi_i, \kappa)$, where ϕ_i is the mean direction of population i and κ its concentration parameter.

Arguments	κ *known*	κ *unknown*
`data`	matrix (one column)	matrix (as many columns as replications)
`levels`	numeric vector with the levels of the order to be contrasted	
`kappa`	numeric value	(NULL)

Table 3: Arguments of the *CTi* function

The three arguments of the function CTi are `data`, `levels` and `kappa`. The characteristics of these arguments are described in Table 3. In this section we explain these arguments with the help of two examples. In the first example, which is based on the data provided Example 1, we assume that κ is known and in the second example, which is based on a set of data available in the package, κ is an unknown parameter.

- **Example 2.1 (κ known):**

  Using the same notation as in Example 1, we test the following hypotheses.

$$H_0: \left\{ \begin{array}{c} \phi_{11} \\ \phi_{12} \\ \phi_{13} \end{array} \right\} \preceq \left\{ \begin{array}{c} \phi_{21} \\ \phi_{22} \end{array} \right\} \preceq \{\phi_{31}\} \preceq \left\{ \begin{array}{c} \phi_{41} \\ \phi_{42} \end{array} \right\} \preceq \left\{ \begin{array}{c} \phi_{11} \\ \phi_{12} \\ \phi_{13} \end{array} \right\}$$

  H_1 : H_0 is not true.

  The `data` argument contains the sample mean directions in the form of a matrix where the rows are the individuals and the columns are the replications when they exist, if not there is one column with the angular means. In this case `data = cbind(cirdata))`. The value of κ is introduced in `kappa`. Thus, for the data and the order restriction in Example 1(page 9), assuming $\kappa = 0.2$ we have the following statements. The output is the p-value for the conditional test from equation (15).

```
> CTi(data = cbind(cirdata), levels = c(1, 1, 1, 2, 2, 3, 4, 4),
+     kappa = 0.2)
```

  Since `p-value=0.9615`, we cannot reject the null hypothesis that the parameters satisfy this general isotropic order.

- **Example 2.2 (κ unknown (replications needed)):**

  Using the data in package called `datareplic` we demonstrate the use of the function *CTi* when κ is unknown. As remarked earlier, when κ is unknown we need replicate data to estimate κ. The file `datareplic` is a matrix with columns denoting replications and rows denoting the angles corresponding to each individual. We have 8 parameters $(\phi_{11}, \phi_{12}, \phi_{13}, \phi_{21}, \phi_{22}, \phi_{31}, \phi_{32}, \phi_{41})$ and the order to be contrast is.

$$H_0: \left\{ \begin{array}{c} \phi_{11} \\ \phi_{12} \\ \phi_{13} \end{array} \right\} \preceq \left\{ \begin{array}{c} \phi_{21} \\ \phi_{22} \end{array} \right\} \preceq \left\{ \begin{array}{c} \phi_{31} \\ \phi_{32} \end{array} \right\} \preceq \{\phi_{41}\} \preceq \left\{ \begin{array}{c} \phi_{11} \\ \phi_{12} \\ \phi_{13} \end{array} \right\}$$

  H_1 : H_0 is not true.

  We take the data from the package. We have to set the levels of the order in the argument `levels`.

```
> data(datareplic)
> orderLevels2 <- c(rep(1, 3), rep(2, 2), rep(3, 2), rep(4, 1))
```

Since replicate data are available, we do not include `kappa` in the *CTi* function. Thus we have the following code:

```
> CTi(datareplic, levels = orderLevels2)
```

The result is the p-value defined in (15). Since the `p-value=0.2660570` we may say that there is not sufficient evidence in the data to reject the null hypotheses that the angles are in an isotropic order.

4. Application to analysis of the cell cycle gene expression data

As commented earlier, the motivation for the development of the methods described in Rueda *et al.* (2009) and Fernandez *et al.* (2011) is the analysis of gene expression data in the cell cycle. In this setting, researchers are interested in identifying and understanding functions of genes participating in a normal cell division cycle in order, for example, to detect disruptions that may lead to excessive proliferation of cells.

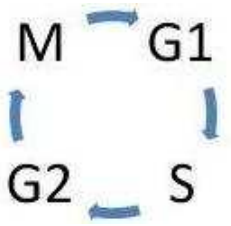

Figure 2: Phases of a cell cycle

A normal cell cycle goes through four phases, shown in the diagram of the Figure 2, Growth 1 (G1), Synthesis (S), Growth 2 (G2) and Mitosis (M). Biologists are interested in determining the phase associated with a cell cycle gene because it may correspond to the biological function of the gene. "The phase associated with a gene" is the phase corresponding to its maximum expression. This moment of peak expression of the gene in the cell cycle is usually called the "phase angle" of the gene.

The length, both of the cycle and the phases, varies a lot depending on the organism. Here, we consider two species of yeasts: *S. Cerevisiae* and *S. Pombe*. They are a good example of the different length of the phases. For instance, there is a great difference between the G2 phase in budding yeast (*S. Cerevisiae*) and in fission yeast (*S. Pombe*), see Figures 3 and 4.

In this example, we consider 16 genes that have a good rank of periodicity in both yeasts in order to ensure the quality of their peak expressions data along the cell cycle. The *S. Pombe* genes, with their corresponding *S. Cerevisiae* orthologs in parentheses are: *ssb1* (RFA1), *cdc22* (RNR1), *msh6* (MSH6), *psm3* (SMC3), *rad21* (MCD1), *cig2* (CLN2), *mik1* (SWE1), *h3.3* (HHT2), *hhf1* (HHF1), *hht3* (HHT1), *hta2* (HTA2), *htb1* (HTB2), *fkh2* (FKH1), *chs2* (CHS2), *sid2* (DBF2) and *slp1* (CDC20).

We test if the order given by the *S. Cerevisiae* genes is maintained by the corresponding *S. Pombe* orthologs. The order for the *S. Cerevisiae* genes is taken from the values given in the

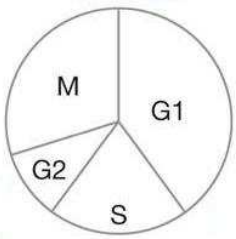

Figure 3: Budding yeast cycle

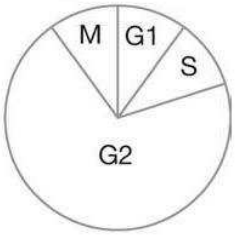

Figure 4: Fission yeast cycle

comprehensive database by Gauthier (2007) available at *http://www.cyclebase.org*. The *S. Pombe* data comes from 10 experiments conducted by different biologists: 3 of them by Oliva *et al.* (2005) , 2 by Peng *et al.* (2005) and 5 by Rustici *et al.* (2004). The unrestricted phase angle values for these *S. Pombe* data have been obtained using the Random Periods Model developed in Liu *et al.* (2004). All these data are in a matrix named `cirgenes` where each column is an experiment and each row is a gene, see Table 4. The genes are ordered according to their corresponding *S. Cerevisiae* orthologs.

We test the *S. Cerevisiae* order in each of the 10 *S. Pombe* experiments. Suppose ϕ_i denotes the phase angle of gene i in *S. pombe* then the hypotheses of interest is:

$$\begin{aligned} H_0 : \phi_{ssb1} \preceq \phi_{cdc22} \preceq \phi_{msh6} \preceq \phi_{psm3} \preceq \phi_{rad21} \preceq \phi_{cig2} \preceq \phi_{mik1} \preceq \phi_{h3.3} \preceq \\ \preceq \phi_{hhf1} \preceq \phi_{hht3} \preceq \phi_{hta2} \preceq \phi_{htb1} \preceq \phi_{fkh2} \preceq \phi_{chs2} \preceq \phi_{sid2} \preceq \phi_{slp1} \preceq \phi_{ssb1} \end{aligned} \tag{18}$$

$H_1 : H_0$ is not true.

We begin with the following code to implement isocir for obtaining CIRE and the SCE values for each of the above 16 genes in the 10 experiments. Results are summarized in Table 5.

```
> data(cirgenes)
> resultIsoCIRE <- matrix(ncol = ncol(cirgenes), nrow = nrow(cirgenes))
> SCEs <- NULL
> for (i in 1:ncol(cirgenes)) {
+     genes <- as.numeric(cirgenes[!is.na(cirgenes[, i]), i])
+     rCIRE <- CIREi(cbind(genes))
+     resultIsoCIRE[!is.na(cirgenes[, i]), i] <- as.vector(rCIRE$CIRE,
+         mode = "numeric")
+     SCEs[i] <- rCIRE$SCE
+ }
```

Now, we use the *CTi* function to perform the conditional test in each of the 10 experiments. Notice that we have no replications here since the experiments were not performed under the same experimental conditions. So for this example we consider κ values obtained from the calculations made in Fernandez *et al.* (2011). The following code gives the p-values for each

experiment using the asymptotic distribution of the conditional test. Results are summarized in Table 5.

```
> kappas <- c(3.958, 3.03, 1.788, 22.475, 14.52, 21.767, 8.607,
+     14.143, 5.945, 14.284)
> pvalues <- NULL
> for (i in 1:ncol(cirgenes)) {
+     genes <- as.numeric(cirgenes[!is.na(cirgenes[, i]), i])
+     k <- kappas[i]
+     pvalues[i] <- CTi(cbind(genes), kappa = k)
+ }
```

From the p-values in Table 5, we see that we cannot reject the null hypothesis that the isotropic order is conserved between the two species of yeasts in any of the 10 experiments. Therefore, it seems plausible that the peak expressions of these 16 genes in *S. Pombe* follow the same order as in *S. Cerevisiae*, which is a very interesting conclusion for evolutionary biologists.

5. Conclusions

In this paper the R package **isocir** has been presented. This package provides useful tools for making inferences for circular data under order restrictions. The first of the two main functions computes CIRE, the circular version of the widely known isotonic regression in R^q. The second one is designed for testing isotropic hypotheses using a conditional test. We have provided the step by step execution of these functions with the **isocir** package. Although we illustrated the proposed methodology using an example from cell biology, the proposed software can be applied to a wide range of contexts. For example, biologists working on circadian clock may be interested in the testing for the conservation of isotropic order among circadian genes between two tissues (e.g. Liu *et al.* (2006)). Similarly, an endocrinologist, studying the order of peak values of various hormones in women during ovulation under different treatment conditions, may find the proposed software useful.

We also want to stress that circular data under restrictions is a field widely open to new developments both in the methods and in implementation. Therefore, it is to be expected that new analysis methods that can lead to new R packages or functions that may appear in the near future.

Acknoledgments

SB's, MAF's and CR's research was partially supported by Spanish MCI grant MTM2009-11161. SDP's research was supported by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences [Z01 ES101744-04].

References

Agostinelli C (2009). ***CircStats**: Circular Statistics, from "Topics in circular Statistics" (2001)*. R package version 0.2-4. URL http://cran.r-project.org/web/packages/CircStats/index.html.

Balabdaoui F, Rufibach K, Santambrogio F (2009). ***OrdMonReg**: Compute least squares estimates of one bounded or two ordered isotonic regression curves*. R package version 1.0.2. URL http://cran.r-project.org/web/packages/OrdMonReg/index.html.

Berens P (2009). "CircStat: A MATLAB Toolbox for Circular Statistics." *Journal of Statistical Softare*, **31**(10), 1–21. URL http://www.jstatsoft.org/v31/i10.

Boles L, Lohmann K (2003). "True Navigation and Magnetic Maps in Spiny Lobsters." *Nature*, **421**, 60–63.

Bowers J, Morton I, Mould G (2000). "Directional Statistics of the Wind and Waves." *Applied Ocean Research*, **22**, 13–30.

Brunsdon C, Corcoran J (2005). "Using Circular Statistics to Analyse Time Patterns in Crime Incidence." *Computers, Environment and Urban Systems*, **30**, 300–319.

Chasalow S (2010). ***combinat**: combinatorics utilities*. R package version 0.0-8. URL http://cran.r-project.org/web/packages/combinat/index.html.

Cochran W, Mouritsen H, Wikelski M (2004). "Migrating Songbirds Recalibrate Their Magnetic Compass Daily from Twilight Cues." *Science*, **304**, 405–408.

Curtis SM (2010). ***bisoreg**: Bayesian Isotonic Regression with Bernstein Polynomials*. R package version 1.0. URL http://cran.r-project.org/web/packages/bisoreg/index.html.

de Leeuw J, Hornik K, Mair P (2011). ***isotone**: Active set and generalized PAVA for isotone optimization*. R package version 1.0-1. URL http://cran.r-project.org/web/packages/isotone/index.html.

Fernandez M, Rueda C, Peddada S (2011). "Isotropic Order among Core Set of Orthologs Conserved between Budding and Fission Yeast." *Preprint*.

Gauthier N (2007). "Cyclebase.org - A Comprehensive Multi-Organism Online Database of Cell-Cycle Experiments." *Nucleic Acids Research*, **36**, 854–859. URL http://www.cyclebase.org/.

Haskey J (1988). "The Relative Orientation of Addresses of Spouses Before Their Marriege: An Analysis of Circular Data." *Journal of Applied Statistics*, **15**, 183.

Jammalamadaka S, SenGupta A (2001). *Topics in Circular Statistics*. World Scientific.

Kibiak T, Jonas C (2007). "Appliying Circular Statistics to the Analysis of Monitoring Data." *European Journal of Psychological Assessment*, **23**, 227–237.

Le C, Liu P, Lindgren B, Daly K, Giebind G (2003). "Some Statistical Methods for Investigating the Date of Birth as a Disease Indicator." *Statistics in Medicine*, **22**, 2127–2135.

Liu D, Peddada S, Li L, Weinberg C (2006). "Phase analysis of circadian-related genes in two tissues." *BMC Bioinformatics*, **7**, 87.

Liu D, Umbach D, Peddada S, Li L, Crockett P, Weinberg C (2004). "A Random Periods Model for Expression of Cell-Cycle Genes." *The National Academy of Sciences of the USA*, **101**(19), 7240–7245.

Lund U, Agostinelli C (2010). ***circular**: Circular Statistics*. R package version 1.1. URL http://cran.r-project.org/web/packages/circular/index.html.

Maldonado P, Godecke I, Gray C, Bonhoeffer T (1997). "Orientation Selectivity in Pinwheel Centers in Cat Striate Cortex." *Science*, **276**, 1551–1555.

Mardia K, Hughes G, Taylor C, Singh H (2008). "A Multivariate von Mises Distribution with Applications to Bioinformatics." *Canadian Journal of Statistics*, **36**, 99–109.

Mardia K, Jupp P (2000). *Directional Statistics*. Wiley.

Morphet C (2009). ***CircSpatial**: Functions For Circular Spatial Data*. R package version 1.0-1. URL http://cran.r-project.org/web/packages/CircSpatial/index.html.

Oliva A, Rosebrock A, Ferrezuelo F, Pyne S, Chen H, Skiena S, Futcher B, Leatherwood J (2005). "The Cell-Cycle-Regulated Genes of Schizosaccharomyces Pombe." *Plos. Biology*, **3**, 1239–1260.

Peddada S, Dinse G, Kissling G (2007). "Incorporating Historical Control Data When Comparing Tumor Incidence Rates." *Journal of the American Statistical Association*, **102**, 1212–1220.

Peng X, Karuturi R, Miller L, Lin K, Jia Y, Kondu P, Wang L, Wong L, Liu E, Balasubramanian M, Liu J (2005). "Identification of Cell Cycle-Regulated Genes in Fission Yeast." *The American Society for Cell Biology*, **16**, 1026–1042.

R Development Core Team (2004). "R: A Language and Environment for Statistical Computing." *R Foundation for Statistcal Computing, Vienna, Austria*. URL http://www.R-project.org/.

Robertson T, Wright F (1980). "Algorithms in Order Restricted Statistical Inference and the Cauchy Mean Value Property." *The Annals of Statistics*, **8**(3), 645–651.

Robertson T, Wright F, Dykstra R (1988). *Order Restricted Statitical Inference*. Wiley.

Rueda C, Fernandez M, Peddada S (2009). "Estimation of Parameters Subject to Order Restrictions on a Circle with Application to Estimation of Phase Angles of Cell-cycle Genes." *Journal of the American Statistical Association*, **104**(485), 338–347.

Rufibach K (2010). ***OrdFacReg**: Least squares, logistic, and Cox-regression with ordered predictors*. R package version 1.0.2. URL http://cran.r-project.org/web/packages/OrdFacReg/index.html.

Rustici G, Mata J, Kivinen K, Lio P, Penkett C, Burns G, Hayles J, Brazma A, Nurse P, Bahler J (2004). "Periodic Gene Expression Program of the Fission Yeast Cell Cycle." *Nature Genetics*, **36**, 809–817.

Turner R (2009). ***Iso**: Functions to perform isotonic regression*. R package version 0.0-8. URL http://cran.r-project.org/web/packages/Iso/index.html.

van Doorn E, Dhruva B, Sreenivasan K, Cassella V (2000). "Statistics of Wind Direction and its Increments." *Physics of Fluids*, **12**, 1529–1534.

Zar J (1999). *Biostatistical Analysis*. Prentice Hall.

Table 4: Initial *S. Pombe* phase angle data for each experiment

	Genes															
Experiments	**ssb1**	**cdc22**	**msh6**	**psm3**	**rad21**	**cig2**	**mik1**	**h3.3**	**hhf1**	**hht3**	**hta2**	**htb1**	**fkh2**	**chs2**	**sid2**	**slp1**
1.Oliva cdc	0.202	0.217	6.261	5.764	0.892	5.611	6.257	1.178	0.911	1.201	0.971	1.288	5.298	5.596	4.251	5.209
2.Oliva elut1	2.939	3.261	2.810	2.848	1.603	2.381	1.709	4.689	4.355	4.717	4.418	4.397	1.601	1.819	1.751	2.518
3.Oliva elut2	0.440	0.447	5.257	6.206	4.381	5.458	6.044	1.541	0.727	6.114	0.351	0.687	3.935	3.970	5.835	5.895
4.Peng cdc	3.327	3.565	3.387	2.806	3.193	3.260	3.026	4.778	4.693	4.755	4.816	4.675	2.685	2.769	2.885	2.421
5.Peng elut	3.333	3.912	3.894	3.443	3.647	3.969	4.296	5.188	5.059	5.143	5.215	5.243	3.338	3.607	3.082	3.185
6.Rust cdc1	1.965	2.151	2.033	2.028	1.741	2.072	1.730	3.129	2.993	3.085	3.063	2.872	1.281	1.178	1.905	1.236
7.Rust cdc2	1.809	2.207	1.414	1.351	1.963	1.940	1.978	3.744	3.584	3.669	3.479	3.590	1.382	1.455	1.063	1.396
8.Rust elut1	—	1.457	1.288	—	1.529	1.373	1.379	2.420	2.278	2.409	2.311	2.245	1.010	1.146	—	1.090
9.Rust elut2	2.213	1.786	1.730	1.987	1.878	1.882	3.071	2.704	2.787	2.814	2.908	2.739	1.351	1.441	1.275	1.420
10.Rust elut3	2.340	2.701	2.703	2.525	2.978	2.319	2.284	3.773	3.567	3.636	3.465	3.431	1.981	1.716	2.523	2.118

Table 5: CIRE, SCE and p-values for each experiment

	CIRE under Isotropic Order																SCE	p-value
Experiments	$\widetilde{\theta}_1$	$\widetilde{\theta}_2$	$\widetilde{\theta}_3$	$\widetilde{\theta}_4$	$\widetilde{\theta}_5$	$\widetilde{\theta}_6$	$\widetilde{\theta}_7$	$\widetilde{\theta}_8$	$\widetilde{\theta}_9$	$\widetilde{\theta}_{10}$	$\widetilde{\theta}_{11}$	$\widetilde{\theta}_{12}$	$\widetilde{\theta}_{13}$	$\widetilde{\theta}_{14}$	$\widetilde{\theta}_{15}$	$\widetilde{\theta}_{16}$		
1.Oliva cdc	6.256	6.256	6.256	6.256	0.054	0.054	0.054	1.045	1.045	1.085	1.085	1.288	5.069	5.069	5.069	5.209	1.269	0.346
2.Oliva elut1	2.526	2.526	2.526	2.526	2.526	2.526	2.526	4.516	4.515	4.515	4.515	4.515	1.600	1.785	1.785	2.518	1.217	0.767
3.Oliva elut2	5.849	5.849	5.849	5.849	5.849	5.849	6.044	0.598	0.598	0.598	0.598	0.687	3.935	3.970	5.835	5.849	2.666	0.299
4.Peng cdc	3.224	3.224	3.224	3.224	3.224	3.224	3.224	4.736	4.736	4.748	4.748	4.748	2.685	2.692	2.692	2.692	0.247	0.432
5.Peng elut	3.333	3.724	3.724	3.724	3.724	3.969	4.296	5.124	5.124	5.143	5.215	5.243	3.302	3.302	3.302	3.302	0.155	0.717
6.Rust cdc1	1.960	1.960	1.960	1.960	1.960	1.960	1.960	3.028	3.028	3.028	3.028	3.028	1.230	1.230	1.571	1.571	0.213	0.679
7.Rust cdc2	1.692	1.692	1.692	1.692	1.951	1.951	1.978	3.613	3.613	3.613	3.613	3.613	1.301	1.301	1.301	1.396	0.296	0.885
8.Rust elut1	—	1.372	1.372	—	1.427	1.427	1.427	2.332	2.332	2.332	2.332	2.332	1.010	1.118	—	1.118	0.028	0.999
9.Rust elut2	1.908	1.908	1.908	1.915	1.915	1.915	2.837	2.837	2.837	2.837	2.837	2.837	1.351	1.358	1.358	1.420	0.124	0.999
10.Rust elut3	2.340	2.585	2.585	2.585	2.585	2.585	2.585	3.574	3.574	3.574	3.574	3.574	1.848	1.848	2.320	2.320	0.268	0.742

Affiliation:

Sandra Barragán, Miguel A. Fernández, Cristina Rueda
Departamento de Estadística e Investigación Operativa
Facultad de Ciencias
Universidad de Valladolid
Valladolid, Spain
E-mail: sandraba@eio.uva.es,
miguelaf@eio.uva.es,
crueda@eio.uva.es
URL: http://www.eio.uva.es/~sandra/
http://www.eio.uva.es/~miguel/
http://www.eio.uva.es/~cristina/

Shyamal D. Peddada
Biostatistics Branch
National Institute of Environmental Health Sciences
Research Triangle Park
NC 27709, USA
E-mail: peddada@niehs.nih.gov,
URL: http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/peddada/index.cfm